DOCUMENT RESUME

ED 088 935 TM 003 496

AUTHOR	Heuer, Edwin
TITLE	Making Standardized Tests Work For Your IGE School.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
CONTRACT	NE-C-00-3-0065
NOTE	8p.
EDRS PRICE	MF-$0.75 HC-$1.50
DESCRIPTORS	Achievement Tests; *Curriculum Evaluation; Educational Improvement; Educational Needs; Elementary School Curriculum; *Evaluation Techniques; Item Analysis; Junior High Schools; Standardized Tests; *Test Validity
IDENTIFIERS	*Individually Guided Education

ABSTRACT

A technique is described which uses the standardized tests to evaluate an Individually Guided Education Curriculum (IGE). A technique was devised so that an evaluation of each test question could be made. The teaching staff at Port Edwards, Wisconsin, where this study was conducted, was asked to rate questions for relevancy to the IGE curriculum and to predict what percent of the class would answer each question correctly. The following scale was used: A, for a valid question; B, for a reasonably valid question; C, for an invalid question. The evaluation was done by grade level teams in grades two through six and by subject matter area in the junicr high, due to departmentalized organization. The data was tabulated and the results appeared to be acceptable but a suggestion was made for the study to be carried further. After a review of the data some valid conclusions were made. Considering only valid and reasonably valid questions, (A or B rating), the tabulations indicated that the students exceeded national percentages at all grade levels. There was an indication that the students did well on items that reflected the IGE curriculum (Author/BB)

MAKING STANDARDIZED TESTS WORK FOR YOUR IGE SCHOOL

A REPORT BY EDWIN HEUER
PRINCIPAL, PORT EDWARDS ELEMENTARY SCHOOL
PORT EDWARDS, WISCONSIN
IN CONJUNCTION WITH
THE WISCONSIN RESEARCH AND DEVELOPMENT
CENTER FOR COGNITIVE LEARNING

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

Wisconsin Research & Development Center, University of Wisconsin, 1025 W. Johnson St.
Madison, Wisconsin 53706

In recent history there has been greater emphasis placed on the need for evaluation of educational programs. Most educators will agree that we need to measure the quality of our product. Disagreement and perhaps, now and then, confusion arise in devising the method and selecting the instruments to be used in accomplishing the analysis.

In our community of Port Edwards, Wisconsin, quality business papers are produced for the world market. It is possible for a paper chemist to determine within minutes the quality of that paper at any point along the production line. By contrast, it is very difficult for our school in Port Edwards to analyze and predict the quality of our educational product: growing children. Of course, there are many elements that make the comparison of these two valuable products impractical, if not impossible. The human element has many variables that make it much more difficult to control, and the amount of time taken to form the products is markedly different, to name only two of the dissimilarities.

In spite of the obstacles, educators have evaluated their programs frequently using standardized achievement tests. This procedure, however, presents another obstacle for an IGE school. The goals and objectives set by IGE schools are frequently more specific than the concepts tested by standardized achievement instruments. Test scores, therefore, at best only provide indirect information about the attainment of an IGE school goal. At worst the test scores provide no such information because the concepts tested are not included in the IGE curriculum.

This project paper reports a technique we used in Port Edwards to obtain information from our achievement tests that was directly related to our IGE curriculum. My purpose is not to report the results of our evaluation. Instead I want to illustrate the technique so that you may adapt it to your situation.

The technique grew out of inputs from several sources. One was partially a result of our participation in the Wisconsin Valley League of Cooperating Schools. The league decided to place the full emphasis of its study facilities into a year's study of assessment, evaluation tools, and techniques. Those matters were foremost in our minds and we were stimulated by several discussions held in this connection. The idea, however, was born the year before when our staff was looking at an updated achievement test. One of the concerns frequently expressed by the staff was that much of the achievement test did not really evaluate the effectiveness of our curriculum. The question in my mind was, "How could we equate the test with the curriculum of our school?"

How We Did It

I believed our staff could identify test items that were not relevant to our curriculum, so I devised a relatively simple instrument on which the staff would evaluate each test question. They were asked to rate questions for relevancy to the curriculum and to predict student success.

Relevancy to curriculum. The staff was asked to rate each test question using the following scale:

A = a valid question - the concept has been taught at that or a previous level
B = a reasonably valid question - it is a concept frequently taught
C = not a valid question - this concept has not been taught at this or a previous level.

Prediction of student success. The staff was asked to make their prediction as to how successful the students would be on each question, i.e., what percent of the class would answer the question correctly?

The form used by the staff had a number, indicating the corresponding question on the achievement test, followed by two short blanks. Of course, each form had a place to identify the level of the test. The staff placed their relevancy rating (A, B, or C) on the first blank and a number indicating the percent of students that they expected would correctly answer the question on the second blank.

Because testing was planned for early fall, the teachers of the previous year were asked to evaluate each test, i.e., the test for 3rd grade level students was rated by the 2nd grade staff. This procedure was used because these staff members knew best whether or not the concepts had been taught.

The staff realized the value of this type of information and proceeded to do a deliberate and thorough analysis of each test. It took more time and effort than I originally had anticipated. The evaluation was done by grade level teams in grades 2 through 6 and by subject matter area in our junior high due to its departmentalized organization.

The number of items to be rated ranged from 179 by the 1st grade team (rating the 2nd grade test) to 534 by the 4th and 5th grade teams (rating the 5th and 6th grade tests, respectively). In general, a three-teacher team rated each test and the ratings given each question were the result of their consensus. I felt that this consensus increased the validity of the rating, but it also increased the amount of time required to complete the task. The time required for each team to complete the rating ranged from one to three hours.

When the forms were completed, I tabulated the results on a form similar to the one used by the teachers but which also included the percent of our students that responded correctly on each given question as well as the national percentage for that question. I also left a margin for making any specific comments. It took me about 25 hours to tabulate the data.

What We Found

When I finished I had a lot of numbers and I felt the results were good, but I felt the study should be carried further. Soon afterwards I

had an opportunity to visit with Don Hubbard, coordinator of evaluation at the Wisconsin Research and Development Center. After reviewing the data I had collected, he suggested several specific areas in which valid conclusions could be drawn.

1. What percentage of the items were rated A? B? C?

Overall, we found that one-half (49%) of the 2,852 items were considered by the staff to be valid, one-third (33%) were identified as reasonably valid, and about one-fifth (18%) were found to be inappropriate.

On a grade by grade basis the test was most relevant for the 6th and 8th grades (80% and 73% "A" ratings). It was the least relevant for the 5th and 7th grades (38% and 24% "C" ratings). On a subject by subject basis the test was most relevant for punctuation, capitals, etc., and the spelling subjects (79% and 75% "A" ratings). The test was least relevant for the science and social studies subjects (31% and 34% "C" ratings).

On an even more detailed basis, we examined the ratings for particular subjects at particular grade levels. There are 17 instances out of 59 in which 80% or more of the items received "A" ratings. These instances occurred mainly in the 6th, 7th, and 8th grades and were concentrated in the reading and language arts subjects. The problem areas were the 5th grade science (58% "C" ratings) and the 7th grade social studies (67% "C" ratings).

With regard to the "B" ratings, the 2nd, 3rd, 4th, and 5th grades had relatively high percentages (between 45% and 68%), suggesting that the test was moderately relevant to the curriculum of those children. In these four grade levels there were ten of 29 instances in which the "B" ratings are more than two-thirds of the total number of ratings for a particular subject. There was only one such instance in the other three grade levels.

In order to obtain the information for this question I prepared a table that listed the subtests down the left-hand margin and the grade levels across the columns. Then I made seven columns for each grade level, one column for the total number of items, three columns for the number of items of each type, and then three columns to contain the percentages.

2. Using both the relevancy ratings and the predictions of student success, does the test appear to be excessively difficult or tricky and conversely does it appear too easy or related to out-of-school learning?

The relevancy ratings and the predictions of success should be closely related: The "A" rated questions should receive high predictions of success and the "C" rated questions should receive low predictions of success. Those cases in which the relationship does not hold provide an indication of a difficult test or an easy test. If a question is rated "A" but with

a low prediction of success then the teachers have identified a difficult or tricky situation. If a question is rated "C" but with a high prediction of success then the teachers have identified an easy situation.

We tabulated the number of "A" rated items for which a student success level of less than 70% was set. Of course, different success levels could be used by other schools. Each item of this type could be evaluated to determine the reason for such a combination, but for our purposes we simply tabulated them by subtest in order to pinpoint possible problem areas. There were very few such areas: Math concepts for 2nd and 7th grades and social studies for 6th and 7th grades were the most notable, having more than 60% of their items included in this tally.

To analyze the converse point of this question, we tabulated the number of "C" rated items for which a student success level greater than 50% was set. This success level is another arbitrary criterion which can be altered by the judgement of other faculties. There were fewer instances of these situations than of the others. The single most notable was 7th grade science, in which 56% of the items were tabulated.

The form I used for this analysis was similar to the one used in the first analysis except that I only needed three columns for each grade: one column for the number of items of a given rating ("A" or "C"), one column for the number of items meeting the criterion, and the other column to contain the percentage.

3. Considering only the valid and reasonably valid questions ("A" or "B" rating), how did our students do when compared to national percentages?

I did this analysis by comparing our percentage correct with the national percentage correct on each item rated "A" or "B". Don suggested that I use a ±5% range around the national percentage to insure that we were "really" above or below the national figure. For example, if the national percentage was 47% correct, then we were above that figure if our percentage was 52% or more and below that figure if our percentage was 42% or below. If our percentage was between 43% and 51%, inclusively, we "broke even".

Our tabulations indicated that we exceeded national percentages at all grade levels, by an amount that ranged from 57% of the items for the 5th grade to 82% of the items for the 6th grade. We broke even (within the ±5% range) on from 14% to 36% of the items and fell below on only 2% to 8% of the items. A modest conclusion on this question was that on relevant and reasonably relevant items we were considerably above national averages. This was also a strong indication that the students did well on items that reflected our curriculum.

More importantly, though, we could pinpoint those areas in which we could be satisfied with our instruction and those areas in which we needed

to evaluate our efforts. There were many instances of the first situation and only a few of the second. For example, in the math computations test for the 5th grade students, 18% of the items had percentages below the national figure. All of these items had "B" relevancy ratings which lessened our concern about the students' performance. In each case, however, the students achieved a lower percentage of correct items than had been predicted by the teachers. If the students had reached the teachers' expectation they would not have been below the national standard. Thus our teachers might have considered modifying their instruction in those areas of relatively low student performance.

The form I used in this analysis was similar to the one used in the first analysis. There were seven columns for each grade, one column for the number of "A" and "B" rated items for each subtest, three columns for the number of items above, below, and "equal to" the national percentage, and three columns for the percentages.

4. Considering the "C" rated questions, how did our students do when compared to national percentages?

This analysis and the form used for it paralleled the one used for the "A" and "B" rated items. Our results indicated that our students exceeded national averages at all grade levels by a range that was from 49% of the items for the 5th grade to 80% of the items for the 8th grade. In the "break even" category, we ranged from 14% to 37% and fell below national percentages from 0% to 14%. It is interesting to see, in these data, that even on items the teachers identified as being invalid, the students did almost as well as with the valid items. One might conclude, if it is not already an accepted fact, that children gain a considerable amount of their knowledge outside the school setting. One could also conclude that it does not make any difference whether or not a concept is included in our curriculum. It does seem that our children are learning concepts in and out of our classrooms at about the same rate.

As in the analysis of the "A" and "B" rated items, not only could we obtain general information about a particular grade level or a particular subject area but we could also identify specific areas for evaluation. For example, in math computation for both the 7th and 8th grades, 37% and 40% of the items respectively, were below the national percentage. Of course, these were "C" rated items and the teachers had also predicted low success levels for them, but we could still ask if the concepts tested here should have been included in our curriculum.

Summary

I have reported a technique by which we used our standardized test to evaluate our IGE curriculum. Although I submitted some of the results to our school board, we are by no means finished. We have identified some questions for further discussion: Should we include the "C" rated items in our curriculum, particularly those with low success levels? Do the

results in language skills-parts of speech indicate weakness in instruction or does it point out a variance between the new English as is presently taught and the more traditional English that is reflected in the achievement test question? What of the "A" and "B" rated items for which we predicted low success . . . are we underestimating our children . . . are our standards too high? Is the area of math computation low as a result of reduced emphasis in the math texts or does it reflect the instructional program? Is there a need for stronger emphasis in the science area at earlier levels of instruction? What of those goals and objectives in our curriculum that were not tested?

Perhaps you see other questions we could pursue. More importantly, however, you may see how you could adapt this technique for evaluating your IGE curriculum. If so, then we have mastered our objective. Good luck.

For more information contact the Evaluation Section of the Wisconsin R & D Center, 1025 W. Johnson, Madison, WI 53706 or Mr. Edwin Heuer, Port Edwards Elementary School, 801 Second Street, Port Edwards, WI 54469.